

# Discovering Local Cooccurring Patterns from Aerial Images

Wenjing Jia, *Student Member, IEEE*, and David Tien, *Senior Member, IEEE*,

**Abstract**—Developing a spatial searching engine to enhance the search capabilities of large spatial repositories for GIS update has attracted more and more attention. Existing methods are usually designed to extract limited types of objects and use only one aspect of features of images. In this paper, we propose to use the local cooccurring patterns to disclose the cooccurring relationships among each dominant local features and use this local cooccurring patterns to recognize an object from aerial images. For this purpose, we investigate three types of local features: colour-based features, texture-based features, and edge-based shape features. In order to facilitate the feature extraction procedure, we first use discontinuity-preserving smoothing methods to filter the input image. Two popular smoothing techniques are tested and compared. Experimental results are presented in this paper.

**Index Terms**—Local cooccurring patterns, object detection, aerial images.

## I. INTRODUCTION

DEVELOPING a spatial searching engine to enhance the search capabilities of large spatial repositories for GIS update, which is able to automate the process of identifying objects from high-resolution aerial images and in turn update the vector data, has now attracted more and more attention of researchers from both the area of data mining and the areas of image analysis and pattern recognition. One of the typical goals of such searching engine is to automatically search the required image for an object and return a list of possible matches for that item. Content-based image retrieval techniques which combine both spatial and textual features of images have been widely used. In a typical spatial searching engine system, users specify one of the objects they are looking for as a basis of the searching; the system searches the required image for the object and returns a list of possible matches for that item. The positions of the matched objects can then be used to update the vector data of GIS system.

The recent interest is mainly in identifying objects from high-resolution aerial images, thanks to the technical development on high-resolution remote sensor. The major interest for such purpose is in identifying small structures, such as individual road segments and buildings. Objects that need to be identified include various man-made buildings, housing, properties, swimming pools, lakes, dams, roads, and vegetation. In terms of road extraction solely, the extraction of roads

in high-resolution images is believed as a more difficult task than detecting from low resolution images [1]. This is because in low resolution aerial images the width of roads typically ranges from one pixel to a few pixels and the extraction of road networks is equivalent to the detection of lines or detection of strong curvilinear structures. However, with high resolution images, the road-width can vary considerably, and additional variations are present such as vehicles, shadows cast by buildings and trees, overpasses, and changes in surface material. These variations often make the extraction of road networks a difficult problem [1]. Much research on automatically extracting roads or buildings from aerial images has been reported during the past ten years. In [2], a comprehensive survey on automatic road extraction for GIS update has been given. Methods for extracting roads from aerial images can be *fully automated* or *semi-automated*. In semi-automated approaches, initial points or directional information of roads are usually needed to be set manually by operators, such as the one presented in [1], where an initial *seed point* is needed from the operator.

The evaluation of the performance of the searching engine has been discussed by some researchers as well. Many use the evaluation method proposed by Heipke et al. in [3]. The quality of road extraction is evaluated via two measures, i.e., *completeness* and *correctness*. The “correctness” denotes the ratio of the length of the correctly extracted roads to the length of all extracted roads, or simply saying, how many detected roads belong to real roads? The “completeness” denotes the ratio of the correctly extracted roads and the length of the reference roads, or in other words, how many roads have been detected? Some researchers, like the authors of [1], also introduced the *accuracy* and *precision* to measure the accuracy of the extraction. In either standards, a *ground truth* is firstly generated manually.

This idea presented in this paper follows a survey work to be appeared in [4]. It has been found that, existing methods are usually designed to extract limited types of objects and use only one aspect of features of images. In this paper, we propose using the local cooccurring patterns [5] to disclose the cooccurring relationships among each dominant local features and use these local cooccurring patterns to recognize objects from aerial images. For this purpose, we investigate three types of local features: colour-based features, texture-based features, and edge-based shape features. In order to facilitate the feature extraction procedure, we first apply image smoothing techniques to smooth the input images. In order not to lose important edge information, discontinuity-preserving smoothing methods are applied to filter the input image. Two

Ms. Wenjing Jia is currently a part-time Research Associate at the Charles Sturt University. She is a full-time PhD student at Faculty of Information Technology, University of Technology, Sydney, PO Box 123, Broadway, NSW, 2007, Australia (wejia@it.uts.edu.au).

Dr. David Tien is currently a Senior Lecturer at School of Information Technology, Charles Sturt University, Panorama Avenue Bathurst, NSW, 2795, Australia (dtien@csu.edu.au).

popular smoothing techniques have been compared. Experimental results are presented in this paper.

In the remaining parts of this paper, the relevant work is firstly summarized and commented in Section II. Then, the basic idea of the local cooccurring patterns is illustrated in Section III. Two popular discontinuity-preserving smoothing techniques are performed and experimental results on aerial images are presented and compared in Section IV. This paper is concluded in Section V. Our future plan to continue this research is also mentioned in this section.

## II. RELEVANT WORK

The idea presented in this paper is triggered by the work reported by Mena and Malpica in [6], where an automatic road extraction method was proposed. The basic procedure of the method is, after a simple noise smoothing process, a texture-analysis-based segmentation operation is performed pixel by pixel to generate a binary-level image which contains “road” pixel candidates and “non-road” pixels. Hence, the crucial step of this method is the *texture-analysis-based segmentation operation*.

The segmentation process is based on texture analysis, which offers a binary image used as the input of a posterior vectorization process. Next, this segmentation method is briefly introduced. The detailed idea of this segmentation method and more segmentation results on other images can be found in three papers, i.e., [7], [8], and [6].

### A. Basic Idea

The image segmentation method proposed by Mena and Malpica is a *supervised* image segmentation method which is based on colour texture analysis. By “supervised” method, it means in order to correctly differentiate the area of interest, a proper sample area of the object is needed as a *training set*. The training set is represented by several pixels under the area of interest. At each pixel, three distances between the studied pixel and the pixels in the training set are measured in three aspects, based on which a final decision is made. By “texture-analysis-based”, it means, in this algorithm texture-related features are defined and extracted.

The idea underlying this segmentation method is that, the simplest way of classifying whether a given pixel in an image, denoted by  $\vec{x}$ , belongs or not to the area of interest is to calculate the *distance* between the pixel and the pixels in the training set, to see if the studied pixel is close, or similar, enough to the training set to be classified as the same group of the training set. Hence, Mena and Malpica used three statistical features to measure the “distance” between a given pixel to the group of pixels in the training set in three aspects. The first distance is used to compare the similarity between the *colour value* of each *individual* pixel to the average colour value of the pixels in the training set. The second distance is used to measure the similarity between the *distribution* of pixels in a small neighbourhood centered at the given pixel and the *distribution* of the pixels in the training set. The third distance is rather complicatedly designed. A multi-dimensional histogram, which is composed of the distributions

of six Haralick features of cooccurrence distribution among the three colour components of a  $3 \times 3$  neighbourhood, is constructed for a small neighbourhood centered at the given pixel. The Bhattacharyya distance is computed between the cooccurrence distribution of the studied neighbourhood and the cooccurrence distribution of the pixels in the training set. When the values of the three distances are obtained, they are fused to yield only one value for each pixel according to the *Dempster-Shafter Theory of Evidence* [9], called *plausibility*. Then, a threshold is manually tuned to binarize the plausibility map into a binary-level image.

### B. Our Comments

The road extraction method as mentioned above is based on a *supervised* colour image segmentation method, as for each studied aerial image (i.e., the target image), from which the object is to be extracted, a “training set” of object profile is needed in advance as the standard template. Although this still does not fully automate the process of image segmentation or object extraction, the idea of the supervised segmentation can be easily applied to other man-made object extraction, such as red roof, blue swimming pool, or other named objects. All the system needs is different training sets for different tasks, which can be specified by the GIS users using simple operations. For instance, the operator can use the mouse to click the interesting objects on the aerial image. The system then automatically picks up a proper area of the clicked object and uses the pixels in the area as the training set. The other procedures can adopt the idea presented in Mena and Malpica’s papers.

Furthermore, it has been noticed from the segmentation results presented in [8] that the selection of the training set may affect the resultant segmentation outcomes. We cite an example from [8] as shown in Figure 1(a) to show this problem. The training set used in [8] for this image is shown in Figure 1(b). The segmentation result using this training set cited from the paper is shown in Figure 1(c). The original figures can be found in [8]. As shown in Figure 1(b), due to the large colour variance in the meadow, the training set has to be set to include both the middle part and the left bottom part to contain as many as possible pixels that have different colour distributions. This is usually not an easy job for normal operators.

Therefore, a certain amount of typical pixels must be included in the training set to have a better performance. This imposes a requirement for the operators which is not preferable. To tackle this problem, we introduce a more automatic selection method, which allows the operators to click any point of an object in order to search the objects which have similar characteristics. This is shown in the area circled by red boundary in Figure 1(d). To achieve this target, not only colour-based features are to be used, but also texture-based features are to be used.

Furthermore, most existing methods for GIS applications only use one type of features of the images for extracting one kind of objects from aerial images. These features vary from one instance to another. Hence, there is not a common way to describe the algorithms if we need a system which is able

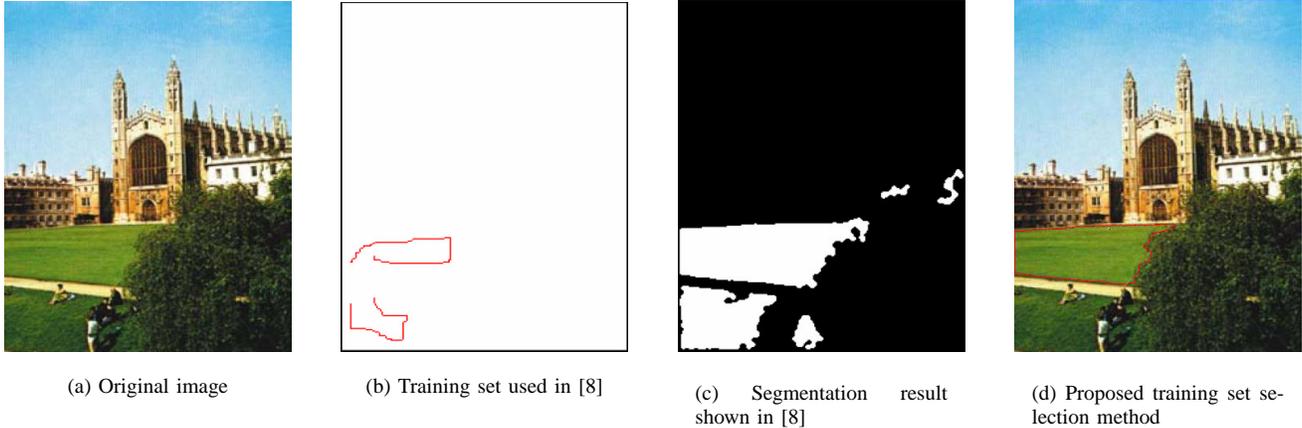


Fig. 1. Selection training set for image segmentation.

to detect more than one kind of objects. The texture-based features introduced in [6] can still be modified in order to improve the overall completeness and the correctness rates. In this paper, we introduce the idea of local cooccurring pattern for this problem. More details are given in next section.

### III. LOCAL COOCCURRING PATTERN

In [5], a novel visual representation method, called *local cooccurring patterns (LCPs)* has been proposed for human detection. The LCPs consist of characteristic local features and the statistical coocurrence relationship between them.

Recognition of visual objects is a fundamental task in vision. Typically, objects to be recognized are represented by many local features or local parts. Test images are processed by extracting local features which are then matched with the object’s model. Most existing work that uses local features assume that each of the local features is independent to each other. However, in many cases, this is not true. We use the following example to illustrate this problem.

Let us consider an example in the following Table I. Assume an image database contains four class of objects, which are “building”, “vegetation”, “roads”, and “grasslands”, with 100 images for each class. Within this image database, two local features, Feature *A* and Feature *B*, are present. Each cell in the table shows the number of images the local features appear in. It can be seen from the table that, both Feature *A* and Feature *B* appear in the four classes evenly. Hence, it is impossible to use either Feature *A* or Feature *B* individually to infer the class label. However, it can also be seen that, Feature *A* and Feature *B* cooccur in the “roads” class much more frequently than the other three classes. Their coocurrence cannot happen by chance. Therefore, there must be an association between these two local features and the “roads” class, a fact that can be used to infer the “roads” class label. This situation also exists when we want to detect other objects.

Hence, the intention of introducing the local cooccurring patterns is to disclose the cooccurring relationships between the local features and use this local cooccurring patterns to recognize an object.

TABLE I

THE COOCCURRENCE OF THE FEATURES IN AERIAL OBJECTS.

Features	Buildings	Vegetation	Roads	Grasslands
Feature <i>A</i>	50	50	50	50
Feature <i>B</i>	50	50	50	50
Feature <i>A&amp;B</i>	0	10	40	5

The idea presented in [6], [7], and [8], uses the colour-based local features only. Hence, it is found to be sensitive to the selection of the training set and may lead to detection failure when there are other areas which have similar colour distribution or when the objects have larger colour variance with the training set. Many existing methods for building extraction use the shape information only. Hence, it is hard to differentiate them from other similar objects which apparently have different colours and edge distributions from the target. For our system, we propose to use the cooccurring patterns of local features, such as colour distribution, texture features, edge-based shape features, for object detection.

### IV. DISCONTINUITY-PRESERVING COLOR IMAGE SMOOTHING

Before the colour features, texture features, and edge-based shape features of images are extracted, it is common that images are smoothed to obtain a smaller colour map and finer edge map by applying a smoothing filter.

Filtering is perhaps the most fundamental operation of image processing and computer vision. In the broadest sense of the term “filtering”, the value of the filtered image at a given location is a function of the values of the input image in a small neighborhood of the same location [10]. For example, Gaussian low-pass filtering computes a weighted average of pixel values in the neighborhood, in which the weights decrease with distance from the neighborhood center increases. This idea indiscriminately blurs the image, removing not only noise but also salient information. Discontinuity-preserving smoothing techniques, on the other hand, adaptively reduce

the amount of smoothing near abrupt changes in the local structure, i.e., edges.

In order to smooth the input image while keeping important edge information as much as possible to enable the extraction of both the color and edge based local features, we study the two most popular *edge preserving smoothing* techniques: bilateral filtering and mean shift filtering.

#### A. Image Smoothing Using Bilateral Filter

The bilateral filtering was introduced by Tomasi and Manduchi in [10] in 1998. Recently, with the extensive usage of the bilateral filtering in other areas than image denoising, such as demosaicking, image abstraction, image retinex, optical flow estimation, etc, a fast algorithm which uses signal processing approach to approximate the standard bilateral filter has been proposed by Paris and Durand in [11].

The basic idea underlying bilateral filtering is to do in the *range* domain of an image what traditional filters do in its *spatial* domain. Two pixels can be *close* to each other, that is, occupy nearby spatial location, or they can be *similar* to one another, that is, have similar values, possibly in a perceptually meaningful fashion [10].

Bilateral filtering replaces the value of a pixel with a weighted average value of those pixels that either have similar values as that of the given pixel or are close to the given pixel. In this form of filtering, a range filter is combined with a domain filter. A domain filter enforces spatial closeness by weighing pixel intensity values with coefficients that fall off as distance of the neighbouring pixel increases [10]. A range filter, on the other hand, assigns greater coefficients to those neighbouring pixel values that are more similar to the given reference pixel value. Hence, the original intensity value at a given pixel would be better preserved after the value replacement, thanks to range filtering. Range filtering by itself is of little use because values of the pixels that are far away from a given pixel should not contribute to the new value. In one word, the kernel coefficients of a bilateral filter are determined by the combined closeness and similarity function. We explain how a bilateral filter works using mathematical terms as follows [10].

Let  $f : R^2 \rightarrow R$  be the original function of an image which maps the coordinates of a pixel, denoted by  $(x, y)$ , to a value in light intensity. Let  $p_0$  be the reference pixel. Then, for any given pixel  $p$  at location  $(x, y)$ , the coefficient assigned to intensity value  $f(p)$  at  $p$  for the range filter is  $k_r(f(p))$  computed by the similarity function  $s$  as:

$$k_r(f(p)) = (f(p), f(p_0)) = e^{-\frac{(f(p)-f(p_0))^2}{2\sigma_s^2}}. \quad (1)$$

Similarly, the coefficient assigned for the spatial domain filter, denoted by  $k_s(p)$  computed by the closeness function  $c$  as:

$$k_s(p) = c(p, p_0) = e^{-\frac{(p-p_0)^2}{2\sigma_s^2}}. \quad (2)$$

Therefore, for the reference pixel  $p_0$ , its new intensity value, denoted by  $h(p_0)$ , is

$$h(p_0) = C^{-1} \sum_{i=0}^{n-1} f(a_i) \times k_s(a_i \times k_r(a_i)). \quad (3)$$

where  $C$  is the normalization constant and is defined as

$$C = \sum_{i=0}^{n-1} k_s(a_i \times k_r(a_i)). \quad (4)$$

Equation 3 above is called a convolution of the image brightness function  $f$  with spatial domain filter  $k_s$  and range filter  $k_r$ . It will take a long time to carry on the convolution processing as shown in 3 if  $n$  is large. Considering about 99.5% of energy is found in the central area of ‘‘Mexico cap’’ (the curve of Gaussian function with parameter  $\sigma$ ) within the radius of  $3\sigma$ , in order to increase the computation speed, Equation 3 in this paper is computed only over a small area (called convolution window) surrounding each reference pixel and covering the disk with center at the reference pixel and radius of  $3\sigma^s$ . In this paper,  $\sigma_s$  is set to be 3.0. The  $\sigma_r$  is experimentally set to be 0.1. With larger  $\sigma_r$ , it has been noticed that the image appears over-blurred.

#### B. Image Smoothing Using Mean Shift Filter

Another edge-preserving smoothing technique which is also based on the same principle, the simultaneous processing of both the spatial and range domains, is mean shift filtering, proposed by Comaniciu and Meer in [12]. It has been noticed that the bilateral filtering uses a static window in both domains. The mean shift window is *dynamic*, moving in the direction of the maximum increase in the density gradient. Therefore, the mean shift filtering has a more powerful adaptation to the local structure of the data.

Let  $\{x_i\}_{i=1,2,\dots,n}$  be the original image points, let  $\{z_i\}_{i=1,2,\dots,n}$  be the points of convergence, and let  $\{L_i\}_{i=1,2,\dots}$  be a set of labels indicating different segmented regions.

- 1) For each image point  $\{x_i\}_{i=1,2,\dots,n}$ , run the mean shift filtering procedure until convergence and store the convergence point in  $z_i = y_{i,c}$ , as shown below:
  - a) For each image point  $\{x_i\}_{i=1,2,\dots,n}$ , initialise  $j = 1$  and  $y_{i,1} = x_i$ . The first subscript  $i$  of  $y_{i,j}$  denotes the  $i$ th image point, and the second subscript  $j$  denotes the  $j$ th iteration.
  - b) Compute  $y_{i,j+1}$  according to Equation 5 until convergence of  $y_{i,c}$ .

$$y_{j+1} = \frac{\sum_{i=1}^n x_i g(\|\frac{x-x_i}{h}\|^2)}{\sum_{i=1}^n g(\|\frac{x-x_i}{h}\|^2)} \quad (5)$$

- c) Assign  $z_i = (x_i^s, y_{i,c}^r)$ , which specifies the filtered data  $z_i$  at the spatial location of  $x_i^s$  to have the range components of the point of convergence  $y_{i,c}^r$ .
- 2) Delineate the clusters (suppose there are  $m$  clusters), denoted by  $\{C_p\}_{p=1,2,\dots,m}$ , in the joint domain by grouping together all  $z_i$  which are closer than  $h_s$  in the spatial domain and  $h_r$  in the range domain under a Euclidean metric, i.e., concatenate the basins of attraction of the corresponding convergence points.
- 3) For each  $i = 1, 2, \dots, n$ , assign  $L_i = \{p | z_i \in C_p\}$ .



Fig. 2. Filtering aerial images using two edge-preserving smoothing methods. The images in the first row are with dimension of  $1024 \times 1024$  pixels. The images in the bottom row are enlarged counterparts of the image areas with dimensions of  $200 \times 270$  pixels.

Before proceeding to describe the algorithms, the issue of which colour space to employ must be settled. To obtain a meaningful segmentation, perceived colour differences should correspond to Euclidean distances in the colour space which has been chosen to represent the features (image points). A Euclidean metric, however, is not guaranteed for all colour spaces. The CIE Luv and CIE Lab colour models are especially designed to best approximate perceptually uniform colour spaces. In both cases, the  $L$ , the lightness (relative brightness), coordinate is defined in the same way. The two spaces differ only through the other two chromaticity coordinates. In our experiments, we employ the CIE Luv colour space.

When applying computer techniques for image processing, an image in two-dimensional (2-D) space can typically be digitised as a 2-D lattice of  $p$ -dimensional ( $p$ -D) vectors, where  $p$  is one for grey-level images and three for colour images to represent the intensity value at the point determined by the 2-D

lattice. The space of the lattice is known as the *spatial domain*, while the domain for the representation of grey level or colour value is known as the *range domain*. For both domains, a Euclidean metric is assumed. After a proper normalisation, the space and range vectors can be concatenated to obtain a joint spatial-range domain of dimension  $d = p + 2$ . For example, both the location and range vectors can be normalised to the range of  $[0, 1]$ . Thus, the multivariate kernel is defined as the product of two radially symmetric kernels, which is defined as follows, and the Euclidean metric allows a single bandwidth parameter for each domain

$$K_{h_s, h_r}(x) = \frac{C}{h_s^2 h_r^2} k\left(\left\|\frac{x_s}{h_s}\right\|^2\right) k\left(\left\|\frac{x_r}{h_r}\right\|^2\right) \quad (6)$$

where  $x_s$  is the spatial part and  $x_r$  is the range part of a feature vector,  $k(x)$  is the common profile used in both domains,  $h_s$  and  $h_r$  are the employed kernel bandwidths, and  $C$  is the corresponding normalisation constant. In our work,

the uniform kernel function is used. We only need to set the bandwidth parameter  $(h_s, h_r)$ , which controls the size of the kernel and this determines the resolution of the mode detection.

Thus, using the mean shift procedure in the joint spatial-range domain, the input colour vehicle images are segmented into many regions. Parameters  $h_r$  and  $h_s$  are also set experimentally.

### C. Experiments on Aerial Image Smoothing

Figure 2(a) shows an example of aerial images. Figures 2(b) and (c) show the filtered image using colour bilateral filter ( $\sigma_s = 3, \sigma_r = 0.1$ ), and mean shift filter ( $(\sigma_s, \sigma_r) = (7, 5)$ , after three iterations) respectively. To see the filtering effects and compare the different filtering effects using the two smoothing techniques, a small image area in each figure has been selected and enlarged, as shown in Figures 2(d), (e) and (f) respectively.

As it can be seen from Figure 2(d) and (f), applying the edge-preserving smoothing techniques, image noise appearing in the original image has been smoothed out. The colour of a building becomes much more uniform. This may lead to a better edge detection result, which in turn leads to a better shape representation, and simpler colour-based feature representation. On the other hand, if we look at the edges, such as the road boundary or the buildings' boundaries, the edges are very well kept and become sharper using either smoothing method.

## V. DISCUSSION AND FUTURE WORK

In the previous sections of this report, we propose using local cooccurring patterns for object detection from aerial images. For this application, we propose to use the cooccurrence patterns of local features, such as colour features, texture features, and edge-based shape features, for object detection purpose. We propose a semi-automatic method to detect any object from aerial images. In this method, the users can use the mouse to click any objects of interest, the systems then automatically computes and picks up an area and uses this as a training set. To facilitate the feature extraction process,

we have studied two popular discontinuity-preserving image smoothing techniques, i.e., bilateral filtering and mean shift filtering, and compared the filtering results on aerial images.

Our future work in the next step is: first, to build feature space on each type of local features; second, to discover the local cooccurring patterns from major local features and use the patterns to train a classifier for object detection.

### ACKNOWLEDGMENT

This project is funded by Cooperative Research Centre for Spatial Information (CRCSI), Australia.

### REFERENCES

- [1] T. Keaton and J. Brokish, "A level set method for the extraction of roads from multispectral imagery," *Proceedings of the 31st Applied Imagery Pattern Recognition Workshop*, pp. 141–147, 2002.
- [2] J. B. Mena, "State of the art on automatic road extraction for gis update: a novel classification," *Pattern Recognition Letters*, vol. 24, no. 16, pp. 3037–3058, 2003.
- [3] C. Heipke, H. Mayr, C. Wiedemann, and O. Jamet, "Evaluation of automatic road extraction," *International Archives of Photogrammetry and Remote Sensing*, vol. 32, pp. 47–56, 1997.
- [4] D. Tien and W. Jia, "Automatic road extraction: a contemporary survey," *Proceedings of the 4th International Conference on Information Technologies and Applications*, p. to appear, 2007.
- [5] H. Wang and P. Miller, "Disvoering the local co-occurring pattern in visual categorization," *Proceedings of the IEEE International Convergence on Advanced Video and Signal based Surveillance Ssystem*, 2006.
- [6] J. Mena and J. Malpica, "An automatic method for road extraction in rural and semi-urban areas starting from high resolution satellite imagery," *Pattern Recognition Letters*, vol. 26, no. 9, pp. 1201–1220, 2005.
- [7] —, "Color image segmentation using the dempster-shafter theory of evidence for the fusion of texture," *Proceedings of the ISPRS Workshop*, vol. XXXIV-3/W8, pp. 139–144, 2003.
- [8] J. B. Mena and J. A. Malpica, "Color image segmentation based on three levels of texture statistical evaluation," *Applied Mathematics and Computation*, vol. 161, no. 1, pp. 1–17, 2005.
- [9] G. Shafer, *A Mathematical Theory of Evidence*. Princeton University Press, 1976.
- [10] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," *Proceedings of the Sixth International Conference on Computer Vision*, pp. 839–846, 1998.
- [11] S. Paris and F. Durand, "A fast approximation of the bilateral filter using a signal processing approach," *Proceedings of the European Conference on Computer Vision (ECCV'06)*, 2006.
- [12] D. Comaniciu and P. Meer, "Mean shift analysis and applications," *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, pp. 1197–1203, 1999.